Published online 14 March 2019

Nucleic Acids Research, 2019, Vol. 47, No. 10 ***e60***
doi: 10.1093/nar/gkz167

DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning

Wenran Li[1,2,3], Wing Hung Wong[2,3,*] and Rui Jiang[1,*]

[1]MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, BNRist, Department of Automation, Tsinghua University, Beijing 100084, China, [2]Department of Statistics, Stanford University, Stanford, CA 94305, USA and [3]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA

Received October 09, 2018; Revised February 08, 2019; Editorial Decision February 27, 2019; Accepted February 28, 2019

Downloaded from https://academic.oup.com/nar/article-abstract/47/10/e60/5380496 by SCD - Universite Toulouse III user on 10 October 2019

ABSTRACT

Interactions between regulatory elements are of crucial importance for the understanding of transcriptional regulation and the interpretation of disease mechanisms. Hi-C technique has been developed for genome-wide detection of chromatin contacts. However, unless extremely deep sequencing is performed on a very large number of input cells, which is technically limited and expensive, current Hi-C experiments do not have high enough resolution to resolve contacts between regulatory elements. Here, we develop DeepTACT, a bootstrapping deep learning model, to integrate genome sequences and chromatin accessibility data for the prediction of chromatin contacts between regulatory elements. DeepTACT can infer not only promoter–enhancer interactions, but also promoter–promoter interactions. In tests based on promoter capture Hi-C data, DeepTACT shows better performance over existing methods. DeepTACT analysis also identifies a class of hub promoters, which are correlated with transcriptional activation across cell lines, enriched in housekeeping genes, functionally related to fundamental biological processes, and capable of reflecting cell similarity. Finally, the utility of chromatin contacts in the study of human diseases is illustrated by the association of *IFNA2* to coronary artery disease via an integrative analysis of GWAS data and interactions predicted by DeepTACT.

INTRODUCTION

Precise identification of physical contacts between regulatory elements is of crucial importance to not only the deciphering of transcriptional regulation, but also the understanding of the mechanisms of human complex diseases. Since human variants that fall into non-coding regions are likely to be responsible for diseases (1), clarifying the effects of functional variants on regulatory elements is key to the understanding of the disease mechanisms. However, most of the non-coding variants are not well annotated and not accurately linked to genes that they regulate (2), making it difficult to evaluate the impact of these mutations. Therefore, precise identification of interactions between promoters and their regulators is urgently needed.

In the past decade, high-throughput methods based on chromosome conformation capture (3C) have been developed to detect physical contacts, but only focus on local loci of the genome (3). Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) technique achieves genome-wide detection, but only captures interactions related to a protein of interest (4). High-throughput chromosome conformation capture (Hi-C), as well as Capture Hi-C, can hopefully realize genome-wide detection of physical chromatin contacts, but requires an extremely deep-sequencing depth to achieve high resolution, which is costly and hard to apply to a large number of cell lines (5,6).

Recently, computational methods have been proposed to improve the resolution of Hi-C data and detect physical interactions at the regulatory element level (7–10). Zhang *et al.* proposed a computational approach, named HiCPlus, to impute the higher resolution interaction maps from low-resolution Hi-C data using a super resolution imaging model (7). Nevertheless, HiCPlus can only improve Hi-C resolution to a level typically not finer than 10 kb, leaving interactions between regulatory elements still unclear. Zhu *et al.* presented EpiTensor, an algorithm to identify 3D spatial associations from 1D maps of histone modifications, chromatin accessibility and RNA-seq data (8). Bkhetan *et al.* developed 3DEpiLoop algorithm to predict chromatin looping interactions from epigenomic data and transcription factor profiles (9). Whalen *et al.* implemented an algorithm called TargetFinder that integrates data for TFs, histone marks, DNase-seq, expression and DNA methylation to predict individual promoter–enhancer interactions across the genome (10). However, all these methods require

*To whom correspondence should be addressed. Tel: +86 135 8198 8092; Email: ruijiang@tsinghua.edu.cn
Correspondence may also be addressed to Wing Hung Wong. Tel: +1 650 725 2915; Fax: +1 650 725 8977; Email: whwong@stanford.edu

© The Author(s) 2019. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

a large number of epigenomic data, which are only simultaneously available in very few human cell lines thus far. Importantly, supervised learning methods like 3DEpiLoop and TargetFinder only focus on the prediction of promoter–enhancer interactions, while recent studies have shown that interactions among promoters are also involved in regulatory processes (11,12). Therefore, a powerful approach to predict genome-wide promoter-related contacts using less epigenomic data is still needed.

Over the past five years, deep neural networks have led to dramatic advances in computer vision and pattern recognition (13,14) and have also been applied to biological problems such as the prediction of DNA accessibility and the recognition of regulatory regions and protein-binding sites (15–17). The success of previous applications of deep neural networks in biological fields inspires us to design a deep learning model to detect chromatin contacts between regulatory elements, utilize the advantage of deep neural networks in automatically learning meaningful feature patterns and capture high-level context dependencies.

In this paper, we develop a bootstrapping deep learning model called DeepTACT (**Deep** neural networks for chromatin con**TACT**s prediction) to predict chromatin contacts at individual regulatory element level using sequence features and chromatin accessibility information. DeepTACT can infer not only promoter–enhancer interactions, but also promoter–promoter interactions. We show that DeepTACT fine-maps chromatin contacts of high-quality promoter capture Hi-C (PCHi-C) from the multiple regulatory element level (5–20 kb) to the individual regulatory element level (1 kb). Besides, DeepTACT identifies a set of hub promoters, which are active across cell lines, enriched in housekeeping genes, closely related to fundamental biological processes and capable of reflecting cell similarity. Moreover, through integrative analysis of chromatin contacts predicted by DeepTACT and existing GWAS data, we inferred novel associations for coronary artery disease, providing a powerful way to build a fine-scale chromatin connectivity map to explore the mechanisms of human diseases.

MATERIALS AND METHODS

Data collection and preprocessing

Promoter capture Hi-C (PCHi-C) data in total B cells (tB), monocytes (Mon), fetal thymus (FoeT), total CD4+ T cells (tCD4), naive CD4+ T cells (nCD4), total CD8+ T cells (tCD8) and 11 other cell types were downloaded from the study conducted by Javierre *et al* (18). In their study, PCHi-C interactions were filtered with a threshold of CHiCAGO scores >5 (19), leaving an average of 25 148 highly confident chromatin loops for each cell type. Processed peaks of DNase-seq data for 199 cell lines were collected from ENCODE (20). Since DNase-seq data are needed in the modeling process, we chose PCHi-C data only in the six cell types that have matching DNase-seq data (i.e. tB, Mon, FoeT, tCD4, nCD4 and tCD8) to train and evaluate our model. Details for the matching data are shown in the Supplementary Data S1. In addition, we collected permissive enhancers from FANTOM5 (21) and extended the length of each enhancer to 2 kb surrounding its middle site, resulting in a list of 65 432 permissive enhancers. We obtained TSS locations from Ensembl release v75 (22) and defined 1 kb regions surrounding TSSs (500 bp upstream and 500 bp downstream) as promoters.

ChIA-PET data in a number of human cell lines were collected from (23) and processed using a standard tool ChIA-PET2 (24) with default settings, yielding 194 467 loops at a q value threshold 0.05. Then, we regarded interactions matched with the loops as validation interactions, yielding 20 504 promoter–promoter interactions and 30 943 promoter–enhancer interactions. Expression quantitative trait loci (eQTLs) were obtained from (25) and were filtered at a q value threshold 0.05. Again, we regarded interactions matched with the eQTLs as validation interactions, yielding 28 144 promoter–promoter interactions and 27 355 promoter–enhancer interactions. Protein–protein interactions (PPIs) were gathered from BIOGRID (26), HPRD (27) and MINT (28) databases, resulting in 74 791 physical interactions in total. Transcripts per kilobase million (TPM) data of four RNA-seq replicates of B cells were collected from ENCODE (20). ChIP-seq profiles of six core histone marks (i.e. H3K4me3, H3K27ac, H3K4me1, H3K4me2, H3K9ac and H3K9me3) and 579 TFs were downloaded from ENCODE (20) on 15 April 15 2017.

Design of DeepTACT model and training strategy

We developed a novel bootstrapping deep learning model, named DeepTACT, to predict chromatin contacts using sequence features and epigenomic information. Specifically, the input for our predictive model is the sequences of two regulatory elements represented with a one-hot encoding strategy (Figure 1A), and their chromatin accessibility scores derived from DNase-seq experiments of a given cell type (Figure 1B). Based on this input, our model will compute the predictive score of whether the two regulatory elements have 3D contact. The model is a deep neural network consisting of three modules: (i) a sequence module for extracting sequence features with two convolutional neural networks (CNNs), (ii) an openness module for learning epigenomic features from chromatin accessibility scores with another two CNNs and (iii) an integration module for merging features of these two modules and gaining higher level context features with an attention-based recurrent neural network (Figure 1C).

In addition, we use an ensemble strategy based on a bootstrapping technique (29) to overcome the instability of the deep neural model caused by random initialization of parameters and local minimum of optimization. First, we bootstrap from an original training set to generate new subsets with the same sample size as the original set. Then, a data augmentation strategy is applied to the resulting subsets to obtain larger datasets for model training. The deep neural network of Figure 1C is trained based on each augmented subset independently, each resulting in a binary classifier. The final output is an ensemble of the binary classifiers derived from different subsets (Figure 1D).

Next, we discuss the important issue of how to train the model. Since our goal is to make context-specific prediction, we need a strategy to construct the training data from the chromatin contact data of that context so that the model can

Downloaded from https://academic.oup.com/nar/article-abstract/47/10/e60/5380496 by SCD - Universite Toulouse III user on 10 October 2019

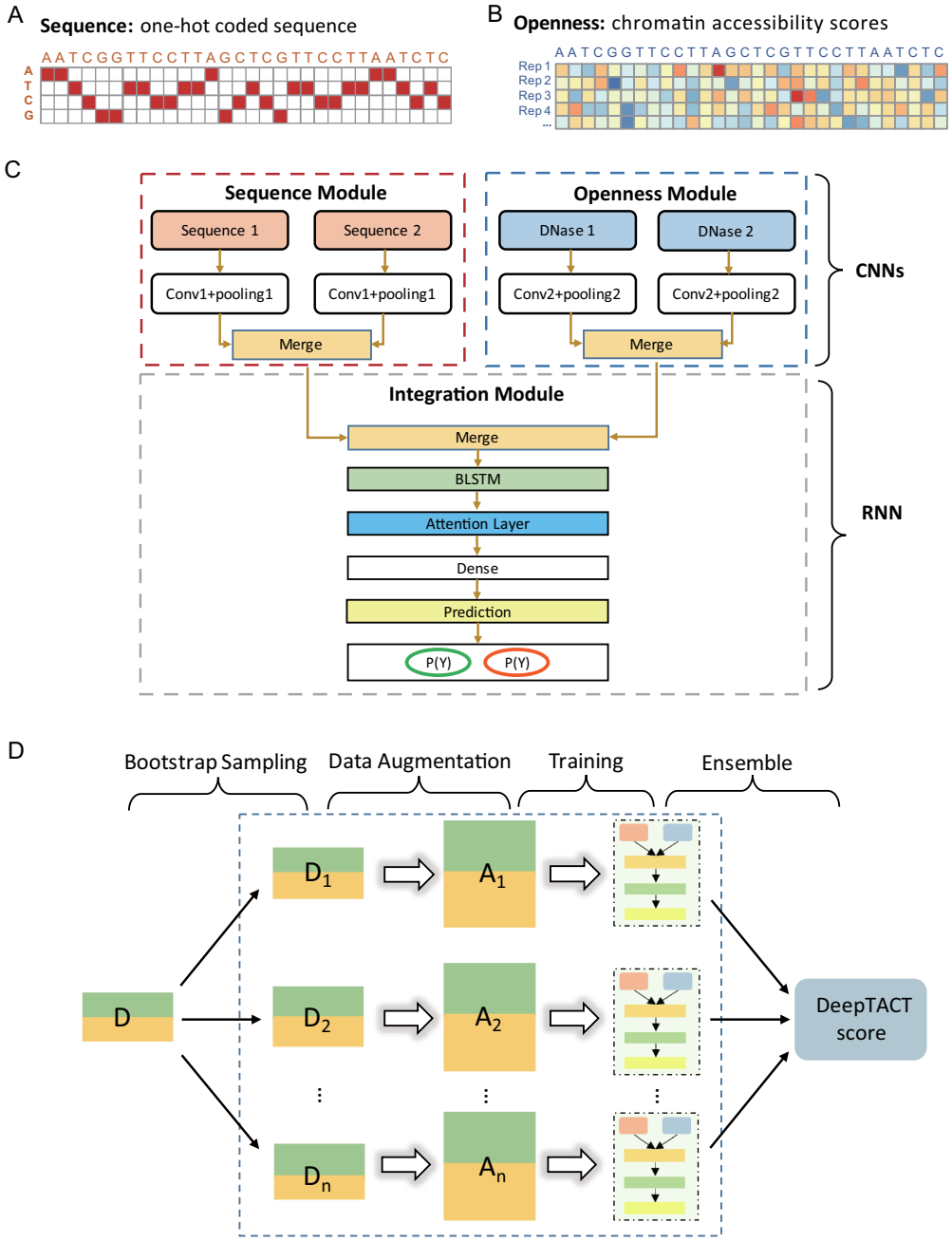

Figure 1. The DeepTACT method. (**A**) One-hot encoded sequence matrix. (**B**) Chromatin accessibility score matrix for replicates of a given cell type. (**C**) Schematic illustration of the deep neural network architecture. (**D**) Schematic illustration of the bootstrapping technique. See 'Materials and methods' section for more details.

be learned from the relevant training data. Suppose we have chromatin contact data in a certain context, for example, promoter capture Hi-C (PCHi-C) data for B cells (18). The data consist of tens of thousands of pairs of interacting regions, where each individual region may be 5–20 kb in size. While a majority of the regions contain multiple regulatory elements, a percentage (about 8.56% in the data in (18)) of the interacting pairs have the property that each region in the pair contains only one regulatory element (Figure 3A and B). Our strategy is to use these pairs, which can indeed identify interacting regulatory elements unambiguously, to construct the positive training cases. Further details on the construction of training data are given in the next section.

Derivation of the chromatin accessibility score for a genomic site

For each DNase experiment, we denote the number of reads falling at each genomic site as N. To remove the effect of sequencing depth, we choose a background window of length W surrounding this peak and denote the number of reads falling into this window as M. The chromatin accessibility

Downloaded from https://academic.oup.com/nar/article-abstract/47/10/e60/5380496 by SCD - Universite Toulouse III user on 10 October 2019

score is formally defined as the fold change of read numbers per base pair and can be simply calculated as

$$O_l = \frac{N}{M/W} \ (l = 1 \ldots L)$$

where L is the length of a given regulatory element. The length of a background window W is set to 1 Mb, according to the suggestion from (30).

Deep neural network

The core structure of the deep neural network used in DeepTACT can be divided into three modules: a sequence module, an openness module and an integration module (Figure 1C). The former two modules are used to learn motif-like patterns from sequences and chromatin accessibility data with separate CNNs. In each CNN, a convolution layer is used for feature extraction, together with a rectifier operation (ReLU) to propagate positive outputs and eliminate negative outputs. Then, a max-pooling layer is used to reduce dimensions and help extract higher level features. In the integration module, the features learned by the above CNNs are concatenated with a merging layer, followed by a bidirectional long- and short-term memory (BLSTM) layer to further learn the context features from the pooled sequence patterns. As a typical representation of recurrent neural networks (RNNs), BLSTM is widely used for its ability to capture dependencies in sequences by accessing long-range context (31). To help the RNN pay more attention to specific sequence patterns, an attention layer (ArXiv: https://arxiv.org/pdf/1512.08756.pdf) is adopted in the integration module, following the BLSTM layer. The final layer of the integration module is a dense layer that is actually an array of hidden units with the ReLU activations feeding into a logistic regression unit that predicts the probability of an interaction. In addition, we adopt batch normalization layers to accelerate the training process and dropout layers to avoid overfitting. Details for the structure and parameters used in the deep learning model are described in Supplementary Table S1.

We implemented the DeepTACT model using Keras 1.2.0 (ArXiv: https://arxiv.org/pdf/1211.5590.pdf) on a Linux server. All experiments were carried out with 4 Nvidia K80 GPUs that significantly accelerated the training process than CPUs.

Bootstrapping strategy

DeepTACT employs a bootstrapping strategy derived from the theory established in (29) for more stable performance (Figure 1D). We first generate K ($K = 20$ in this paper) new datasets of equal size as the original training set by random sampling with replacement from the training data. Then, we apply a data augmentation strategy to each new dataset D_i, yielding an augmented dataset A_i. After that, a deep neural network as described above is trained based on each augmented dataset independently, resulting in an ensemble of the binary classifier $\{S_i\}$. Given the information of a sample as input, its final prediction score is the average of the prediction scores derived from all classifiers, as

$$\text{Score} = \frac{1}{K} \sum_{i=1}^{K} S_i.$$

Data augmentation

Since training a deep learning model needs a large amount of data, we augmented original training sets 20-fold for more stable parameters and better performance. For each element of a positive training sample, we scanned a certain region (say, 2 kb) surrounding the center of the element with a 1 kb sliding window at a step size of 50 bp, yielding 20 substitutions for the element and thus 400 substitution pairs (20×20) for the positive sample. We augmented each positive sample by randomly selecting 20 interactions from its substitution pairs. As for augmentation of negative sets, we simply generated 20 times more negative samples based on the distance distribution of original positive samples.

Activity of hub promoters

For each cell line, we define an activity score for a peak by calculating the fold change between the number of reads falling into a peak and the number of reads falling into a background region surrounding the peak (say, 1 Mb). If the maximum activity score of peaks overlapping a promoter is more than 1, we consider this promoter is active. For each hub promoter, we utilize the number of cell lines where it is active to assess its activity.

Prioritization of disease-related regulatory elements with a gene–enhancer network

We developed a random walk strategy to score the association between genes/enhancers and a given disease by simulating a process that a walker randomly wonders on a gene–enhancer network with certain start probabilities derived from GWAS data. First, we construct a gene-enhancer network using interactions predicted by DeepTACT. Then, we convert the GWAS summary statistics from the SNP level to the promoter/enhancer level using a tool PASCAL (32), yielding a P-value for each node in the network. After that, we transfer the P values of nodes into initial probabilities for random walking, which reflects our prior knowledge about the relationship between the genes/enhancers and the given disease. During the journey, the walker may choose to restart from a new node with the probability π, or to continue the current journey and jump to one neighbor of the current node with the probability $(1 - \pi)$. After a number of steps, the probability that the walker stays at each node would be stable. The steady-state probability of each node can be considered as the association between this node and the given disease, with the information of the gene–enhancer network incorporated.

In mathematical term, we derive the initial probability for an element e (i.e. a gene or an enhancer) with the P-value p_e as

$$p_e^{(0)} = \frac{\exp(\alpha + \beta |\log p_e|)}{\sum_{e=1}^{n} \exp(\alpha + \beta |\log p_e|)},$$

Downloaded from https://academic.oup.com/nar/article-abstract/47/10/e60/5380496 by SCD - Universite Toulouse III user on 10 October 2019

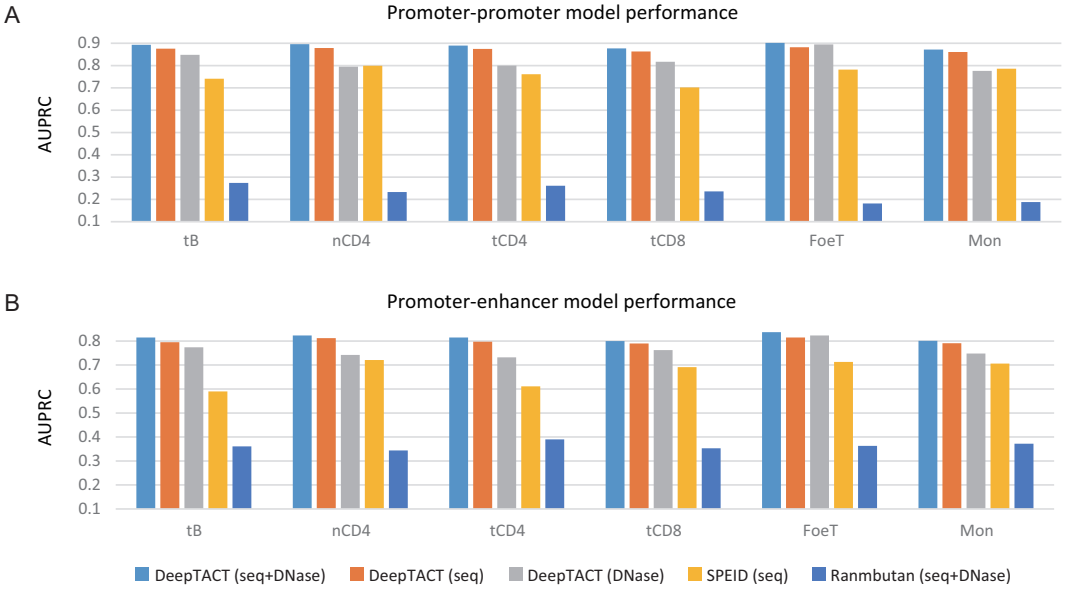

Figure 2. Performance evaluation of DeepTACT. Model comparison of promoter–promoter interactions (**A**) and promoter–enhancer interactions (**B**) in six cell types.

where α and β are tuning parameters ($\alpha = 0$ and $\beta = 1$), n the total number of nodes in the gene–enhancer network. Thus, we derive the initial probabilities as the vector $\mathbf{p}^{(0)} = (p_1^{(0)}, \cdots, p_n^{(0)})$. The gene–enhancer network is represented by a connecting matrix $\mathbf{W} = (w_{gh})_{n\times n}$, where $w_{gh} = 1$ when there is an interaction between node g and node h, otherwise $w_{gh} = 0$. We derive a transition matrix $\mathbf{T} = (t_{gh})_{n\times n}$ by applying row normalization to $\mathbf{W}$, as $t_{gh} = w_{gh} / \sum_{h=1}^{n} w_{gh}$. Then, we recursively update $\mathbf{p}^{(t+1)}$, the probability of staying at each node for time $(t+1)$, with

$$\mathbf{p}^{(t+1)} = (1-\pi)\ \mathbf{T}^T\mathbf{p}^{(t)} + \pi\mathbf{p}^{(0)}.$$

Repeating the iteration a number of steps until $\mathbf{p}^{(t)}$ is stable (e.g. $\Delta p = \|\mathbf{p}^{(t+1)} - \mathbf{p}^{(t)}\| < 10^{-3}$), we obtain the steady-state probability $\mathbf{p}^{(\infty)}$. Alternatively, we derive the probability of the steady state with

$$\mathbf{p}^{(\infty)} = (1-\pi)\ \mathbf{T}^T\mathbf{p}^{(\infty)} + \pi\mathbf{p}^{(0)}.$$

Solving this linear equation gives us the closed form of $\mathbf{p}^{(\infty)}$, as

$$\mathbf{p}^{(\infty)} = \pi\left(\mathbf{I} - (1-\pi)\,\mathbf{T}^T\right)^{-1}\mathbf{p}^{(0)}.$$

It has been shown that results derived from the closed form are consistent with that derived from the simulation process (33). As defaults, we set $\pi = 0.5$. We found that the results of the random walk were robust to the value of π (Supplementary Table S2), which is in accord with the conclusion of previous studies that the random walk model is not sensitive to the parameter π (34).

RESULTS

DeepTACT accurately predicts chromatin contacts

We designed a series of experiments to systematically evaluate the ability of DeepTACT in capturing promoter–promoter interactions and promoter–enhancer interactions. Taking data processing for promoter–promoter interactions as an example, we first regarded unambiguous promoter–promoter interactions as positive samples and uniformly divided them into 10 subsets: one for testing, and the others for training. Then, we generated negative samples under the constraint that the distance between two promoters has the same distribution as that of positive samples (Supplementary Text S1). For training sets, the number of negative samples and the number of positive samples are the same. For testing sets, we generated negative samples five times of positive ones (Supplementary Figure S1 explains the rationale of this imbalance in the test data). Data processing for promoter–enhancer interactions is performed in a similar way.

For the six datasets described in 'Materials and methods' section (Supplementary Table S3), DeepTACT yields AUPRCs of 0.87–0.90 and AUROCs of 0.96–0.97 for promoter–promoter interactions, and AUPRCs of 0.80–0.84 and AUROCs of 0.93–0.95 for promoter–enhancer interactions (Figure 2 and Supplementary Tables S4–S7). We found that DeepTACT with both sequences and chromatin accessibility data as input outperformed the simplified DeepTACT model with either sequences or chromatin accessibility data as input (Figure 2 and Supplementary Tables S4–S7). This indicates that chromatin accessibility data provide complementary information to sequences in detecting cell-type specific interactions. In addition, we collected transcription factors (TFs) reported to be most closely related to each cell type (35), and we found these key TFs were captured by at least two ensemble models of DeepTACT, indicating that the sequence patterns learned by DeepTACT are informative (Supplementary Text S2 and Table S8). Finally, we compared DeepTACT with other state-of-art methods: SPEID (BioRxiv: https://doi.org/10.1101/085241) and Rambutan (BioRxiv: https://doi.org/10.1101/103614)

Downloaded from https://academic.oup.com/nar/article-abstract/47/10/e60/5380496 by SCD - Universite Toulouse III user on 10 October 2019

(details in Supplementary Text S3). With the same testing sets, DeepTACT achieves a mean AUPRC score of 0.89 for promoter–promoter interactions compared with 0.76 of SPEID and 0.23 of Rambutan. For promoter–enhancer interactions, DeepTACT achieves a mean AUPRC of 0.82 compared with 0.67 of SPEID and 0.36 of Rambutan (Figure 2 and Supplementary Tables S4–S7). To stay unbiased, we further evaluated all methods in a dataset of CD34 cell line (6), which was not used for model construction in either method. In this new dataset, DeepTACT still achieves the best performance against other methods, indicating that the superior performance of DeepTACT is robust to PCHi-C datasets (Supplementary Figure S2).

Taken together, the above results show that DeepTACT is capable of integrating sequences and chromatin accessibility data together to identify chromatin contacts between regulatory elements.

DeepTACT provides finer mapping of promoter–promoter interactions from promoter capture Hi-C data

PCHi-C technique identifies pairs of interacting regions, where the length of each region depends on the data resolution. In the above, we discussed how to train and test the DeepTACT model using interacting regions with each end containing only one regulatory element. Once the model has been learned in this way, we can apply it to infer contacts between regulatory elements in situations where one or both interaction regions contain multiple regulatory elements (Figure 3C). We test the performance of this inference on a PCHi-C dataset of B cells at the resolution of 15 kb, to check whether our model can predict element-level interactions from PCHi-C data. We first collected all candidate promoter–promoter (P–P) interactions from the dataset, where a candidate P–P interaction is a possible interaction between two promoters, one from each of the two interaction regions. Then, we used the model trained in B cells to detect true interactions from all candidate pairs. To guarantee prediction precision, for each pair of interacting regions, we predicted interaction only for the pair of promoters with the highest DeepTACT score among the set of candidate P–P interactions. We compared the co-opening of predicted interaction pairs with that of other candidate interaction pairs and found that two promoters of a predicted interaction pair are more likely to be co-opening (Supplementary Text S4 and Figure S3A). This explains why openness data can contribute to the prediction of chromatin contacts. In the B-cell data, the prediction gives an interaction group of 14 691 promoter–promoter interactions. We call this set of interactions the DeepTACT P–P group.

Then, we generated a candidate P–P group by random sampling from all candidate P–P interactions, where the size of the random sample is the same as that of the DeepTACT P–P group (Supplementary Text S5). We also constructed a co-opening control group that selected significant co-opening interactions from candidate P–P interactions (Supplementary Text S5). In addition, we generated a random control group by sampling random P–P pairs from the whole genome, where the distance between two promoters of an interaction pair is the same as that in the DeepTACT P–P group. We checked the overlaps between interactions derived from ChIA-PET data (23) and each P–P group. The DeepTACT P–P group was found to be supported by ChIA-PET data significantly more often than the other three groups (Figure 3D; P values $< 2.2 \times 10^{-16}$, one-sided Wilcoxon rank-sum tests). Moreover, we regarded eQTLs (25) and PPIs (26–28) as additional validation datasets and checked their overlaps with different interaction groups. Again, DeepTACT P–P group was significantly better validated than the other three groups (Figure 3D; P values $< 2.2 \times 10^{-16}$, one-sided Wilcoxon rank-sum tests), indicating that the interactions inferred by DeepTACT were more biologically meaningful than the comparison groups. Meanwhile, results show that in all validation databases, interactions predicted by DeepTACT perform significantly better than interactions selected based on co-opening degrees, which in turn show better performance than original candidate interactions derived from PCHi-C data. This indicates that DeepTACT does not simply predict interactions between regulatory elements based on their chromatin openness.

Next, we asked whether promoters of an inferred interaction pair were more likely to be functionally related than those of a co-opening interaction or a candidate interaction. To answer this question, for two promoters of each interaction pair, we checked their co-occurrence in KEGG pathways (36), REACTOME pathways (37) and GO terms (38) (Supplementary Text S6). In all these databases, the co-occurrence frequency of promoters in the DeepTACT P–P group was significantly higher than those in the co-opening P–P group or the candidate P–P group (Figure 3E; P values $< 2.2 \times 10^{-16}$, one-sided Wilcoxon rank-sum tests), indicating that the DeepTACT-inferred interactions tend to connect functionally related genes more often.

Recently, Dao *et al.* identified mammalian enhancer-like promoters (epromoter) with distal enhancer functions. These 'epromoters' tend to have ubiquitous activity across cell types (39). We collected 493 epromoters in HeLa cells and 632 epromoters in K562 cells (39) and used the 146 epromoters detected in both cell lines for following analysis. We compared the number of epromoters involved in DeepTACT P–P group, co-opening P–P group, candidate P–P group and random P–P group. The result shows that interactions predicted by DeepTACT have significantly the largest overlap with epromoters (Supplementary Figure S4A; P values $< 2.2 \times 10^{-16}$, one-sided Wilcoxon rank-sum tests). We also compared the expression level of genes regulated by epromoters defined by different P–P groups and found that genes regulated by epromoters in DeepTACT group show significantly the highest expression level (Supplementary Figure S4B; P values $< 2.2 \times 10^{-16}$, one-sided Wilcoxon rank-sum tests). These results indicate that interactions predicted by DeepTACT are more biologically meaningful than those derived from PCHi-C data.

DeepTACT provides finer mapping of promoter–enhancer interactions from promoter capture Hi-C data

Similarly, we applied DeepTACT to identify true interactions from all candidate promoter–enhancer (P–E) pairs derived from the PCHi-C dataset of B cells, yielding a DeepTACT P–E group of 8960 promoter–enhancer interactions.

Downloaded from https://academic.oup.com/nar/article-abstract/47/10/e60/5380496 by SCD - Universite Toulouse III user on 10 October 2019

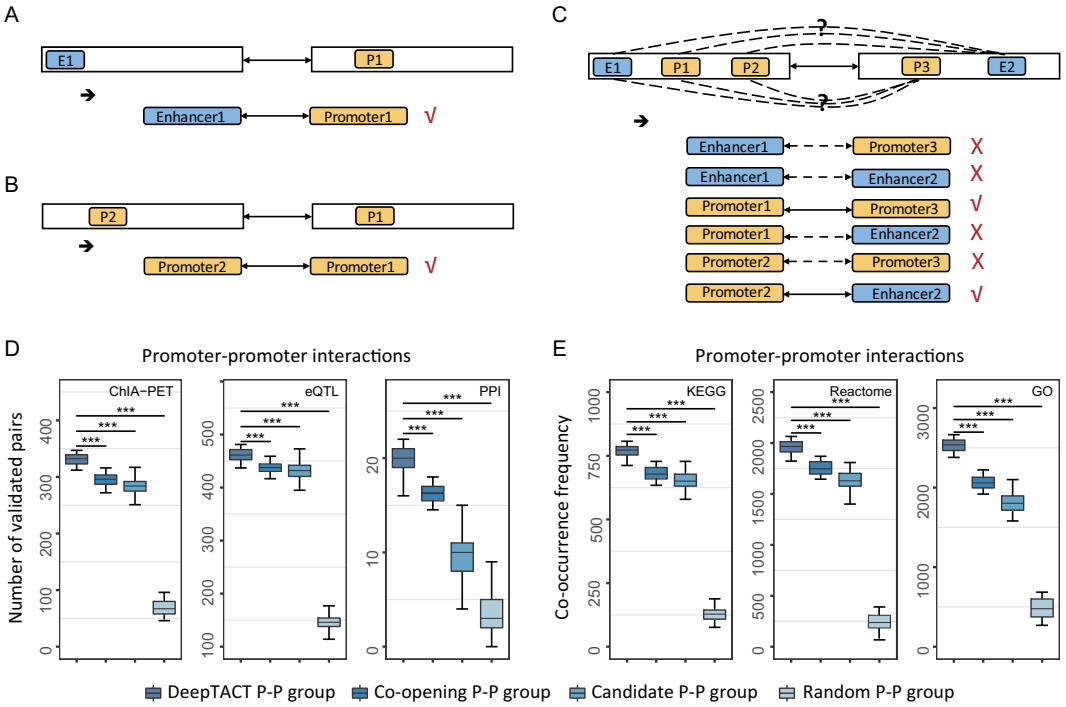

Figure 3. DeepTACT provides finer mapping of promoter–promoter interactions from PCHi-C data. (**A** and **B**) A pair of interacting regions with only one regulatory element in each region. (**C**) A pair of interacting regions with one or both regions containing multiple regulatory elements, resulting in several candidate promoter-level interactions. (**D**) Comparison of DeepTACT P–P group, co-opening P–P group, candidate P–P group, and random P–P group in terms of interactions validated by ChIA-PET data, eQTLs and PPIs. (**E**) Comparison in terms of co-occurrence frequencies of interaction pairs in KEGG pathways, REACTOME pathways and GO terms. For each type of interactions, 1000 subgroups were sampled at the same sample size to generate a distribution (details in Supplementary Text S5). *P* values are based on one-sided Wilcoxon rank-sum tests. *** indicates $P < 2.2 \times 10^{-16}$.

Again, we found that the promoter and enhancer of a predicted interaction pair are more likely to be co-opening than those of other candidate interaction pairs (Supplementary Figure S3B). As in the analysis of P–P interactions, we generated a co-opening P–E group, a candidate P–E group and a random P–E group as controls. We calculated the overlaps between interactions of each group and interactions derived from ChIA-PET data or eQTLs. We found that interactions predicted by DeepTACT were supported by the databases significantly more often than interactions of the other groups (Figure 4A; *P* values $< 2.2 \times 10^{-16}$, one-sided Wilcoxon rank-sum tests), this again indicating that inferred interactions were more biologically significant than interactions derived based on chromatin accessibility or directly derived from PCHi-C data.

Studies have shown that genes regulated by distal enhancers tend to have higher expression level (40). Here, we asked whether genes with distal enhancers defined by interactions in the DeepTACT P–E group tended to have higher expression levels than those defined by the co-opening P–E group or the candidate P–E group. We collected four RNA-seq replicates of B cells from ENCODE (41), and used transcripts per million (TPM) to value gene expression level. We compared the expression level of regulated genes defined by different P–E groups. As shown in Figure 4B, regulated genes defined by DeepTACT tend to have significantly higher expression level than those defined by co-opening P–E interactions or candidate interactions (*P* values < 0.005, one-sided Wilcoxon rank-sum tests), indicating promoter–enhancer interactions inferred by DeepTACT are more related to gene expression than those derived based on chromatin accessibility or directly derived from PCHi-C data.

We further checked the functional enrichment of genes regulated by distal enhancers in B cells and found these genes tend to be enriched for functions related to metabolic processes (Supplementary Figure S5A), which is consistent with previous findings (42,43). In contrast, genes without any distal enhancer did not show significant enrichment in these processes. In addition, we compared the functional enrichment of regulated genes defined by different P–E groups and found the most significant enrichment level in those genes defined by the DeepTACT P–E group (Figure 4C). More specifically, in 8 out of top 10 enriched GO terms, regulated genes defined by DeepTACT demonstrate higher enrichment than those defined by the candidate P–E group (Supplementary Figure S5A). Noticing that genes regulated by distal enhancers showed significant enrichment in key GO functions while those without distal enhancers did not, we developed the following strategy to roughly annotate new functions for enhancers. For a cluster of genes showing enrichment in a certain GO term, we annotate this GO function to the enhancer cluster that regulates those genes (Supplementary Figure S5B). Annotation results are shown in Supplementary Data S2.

Collectively, these results indicate that interactions predicted by DeepTACT have stronger biological meaning than interactions selected based on chromatin accessibility and interactions directly derived from PCHi-C data. To

Downloaded from https://academic.oup.com/nar/article-abstract/47/10/e60/5380496 by SCD - Universite Toulouse III user on 10 October 2019

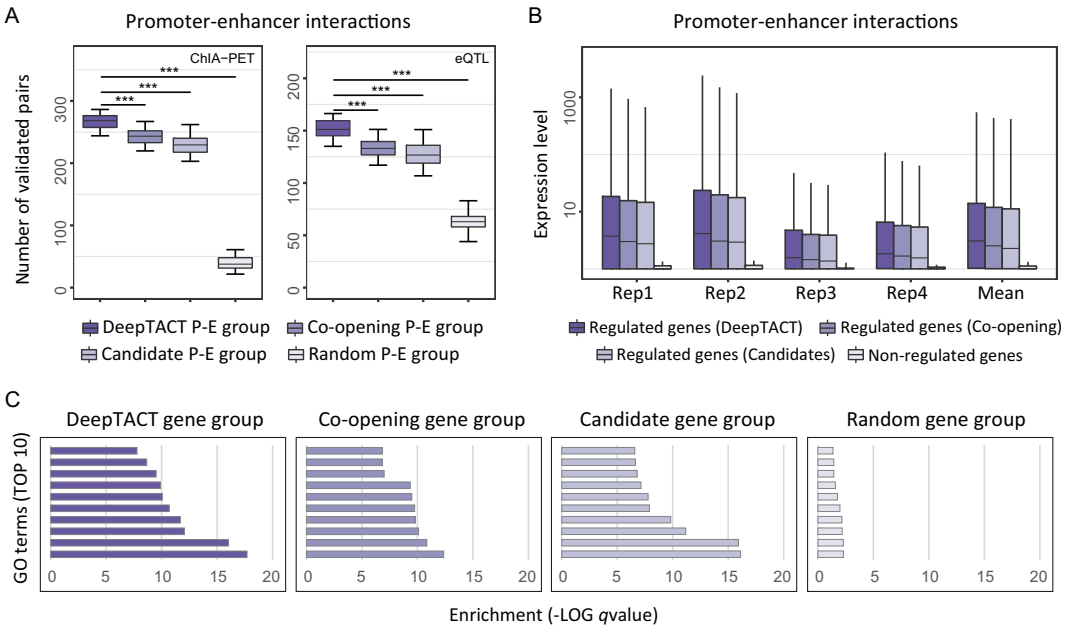

Figure 4. DeepTACT provides finer mapping promoter–enhancer interactions from PCHi-C data. (**A**) Comparison of DeepTACT P–E group, co-opening P–E group, candidate P–E group and random P–E group in terms of interactions validated by ChIA-PET data and eQTLs. (**B**) Comparison of expression level of regulated genes defined by DeepTACT interactions, regulated genes defined by co-opening interactions, regulated genes defined by candidate PCHi-C interactions and non-regulated genes. For each type of interactions, 1000 subgroups were sampled at the same sample size to generate the distribution (details in Supplementary Text S5). *P* values are based on one-sided Wilcoxon rank-sum tests. *** indicates $P < 2.2 \times 10^{-16}$. (**C**) Top 10 enriched GO terms of regulated genes defined by DeepTACT interactions, regulated genes defined by co-opening interactions, regulated genes defined by candidate PCHi-C interactions and randomly selected non-regulated genes.

demonstrate the general applicability of DeepTACT model, we further applied the whole training and prediction process to a new dataset of GM12878 cell line derived from Carins *et al.* (19) and examined the predicted interactions with high quality Hi-C data from Rao *et al.* (5). The result shows that interactions predicted by DeepTACT have significantly more overlaps with Hi-C contacts than other candidate interactions (Supplementary Text S7 and Figure S6), which indicate that our DeepTACT model is generally applicable to PCHi-C datasets.

Characterization of hub promoters defined by predicted interactions

Previous studies have shown that there is a small portion of regulatory elements that tend to be involved in significantly more interactions (8,44). We found that, indeed, interactions were not uniformly distributed among promoters across cell lines (Supplementary Figure S7). Taking the analysis of B cells as an example, we defined the top 10% promoters most frequently involved in chromatin contacts as hub promoters, yielding 1302 hub promoters in B cells. We examined multiple genomic signals to assess the characterization of these hub promoters.

First, we collected 1256 ChIP-seq profiles of six core histone marks (H3K4me3, H3K27ac, H3K4me1, H3K4me2, H3K9ac and H3K9me3) from ENCODE (41) (Supplementary Table S9). For each histone mark, we checked the activity of a hub promoter by counting the number of cell lines where the promoter is active. For comparison, we also extracted promoters with the highest interaction degrees defined by candidate interactions. Excluding hub promoters, we also randomly generated non-hub promoters as a control group. All promoter groups are at the same sample size. As shown in Figure 5A, for histone marks enriched in transcriptionally active promoters (45–47) (i.e. H3K4me3, H3K27ac, H3K4me1, H3K4me2 and H3K9ac), hub promoters defined by DeepTACT interactions were significantly more active across cell lines than the other two promoter groups. As for H3K9me3, which is related to transcriptional repression (48), hub promoters defined by DeepTACT showed the lowest activity across cell lines (Figure 5A; *P*-value < 0.05, one-sided Wilcoxon rank-sum test). Altogether, these results indicate that hub promoters defined by DeepTACT are related to transcriptional activation across cell lines.

Second, we checked the activity of hub promoters in 4383 ChIP-seq profiles of 579 TFs collected from ENCODE (41). For each promoter, we counted the number of cell types where the promoter was active and calculated the total number of covered TFs. We found that hub promoters defined by DeepTACT interactions were significantly more active across cell lines and covered more TFs than those defined by candidate interactions (Figure 5B; *P* values < 0.005, one-sided Wilcoxon rank-sum tests). The comparison result is another supportive evidence for the effectiveness of our DeepTACT model.

Third, based on the finding that hub promoters were active across cell lines, we asked whether hub promoters are enriched in housekeeping genes, which are known to be required for the maintenance of basic cellular function and are expressed in most cells (49). We collected 3669 house-

Downloaded from https://academic.oup.com/nar/article-abstract/47/10/e60/5380496 by SCD - Universite Toulouse III user on 10 October 2019

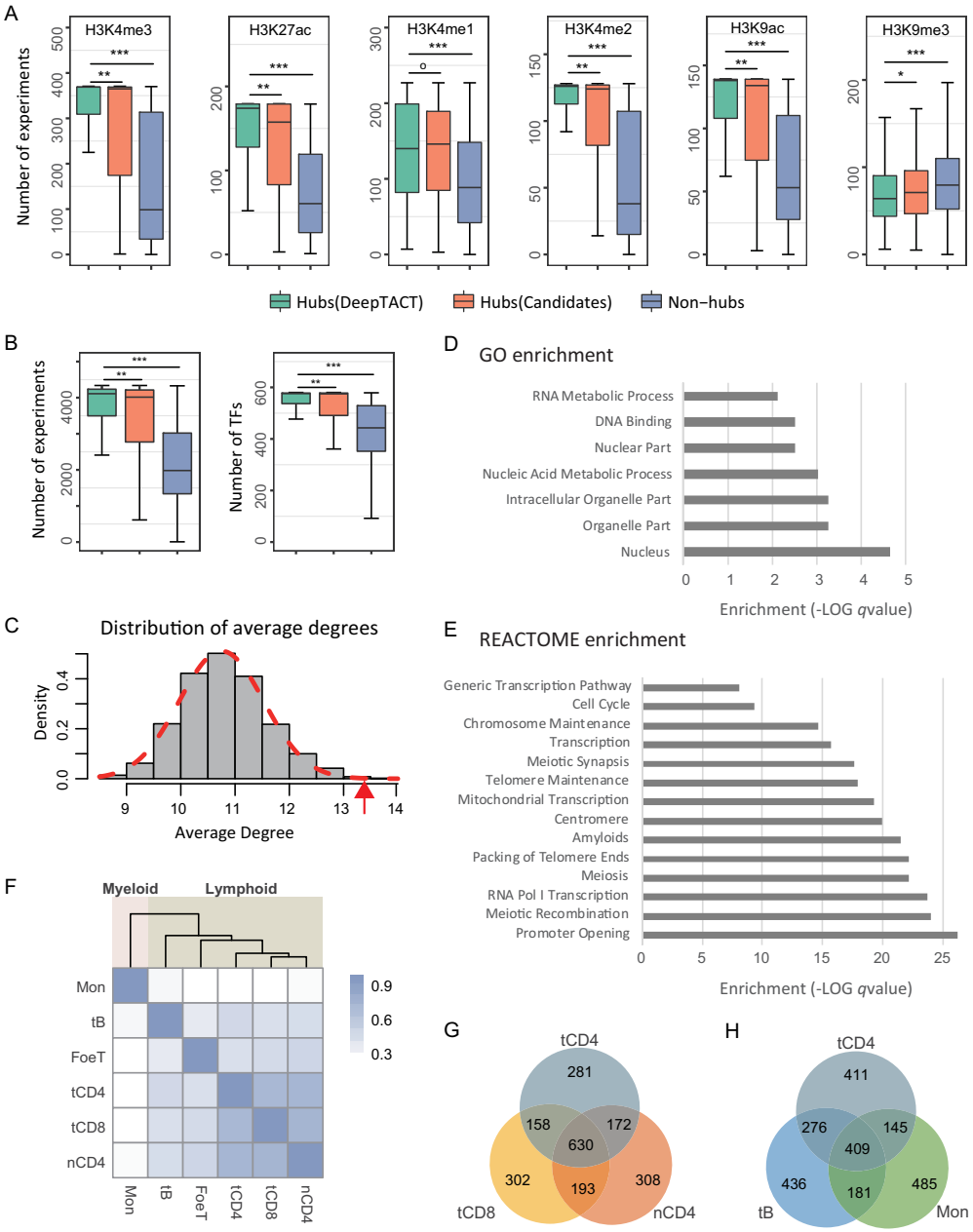

Figure 5. Characterization of hub promoters. (**A** and **B**) Comparison of hub promoters defined by DeepTACT interactions, hub promoters defined by candidate interactions and non-hub promoters in terms of (A) six core histone marks and (B) TFs. The *y*-axis represents the number of experiments where a promoter is active. *P* values are based on one-sided Wilcoxon rank-sum tests. *** indicates $P < 2.2 \times 10^{-16}$, ** $P < 0.005$, * $P < 0.05$ and ^{o} $P > 0.05$. (**C**) Distribution of the average degree of promoter groups in a PPI network. The red arrow represents the average degree of hub promoters defined by DeepTACT interactions. (**D** and **E**) Top enriched GO terms and REACTOME pathways. (**F**) Hierarchical clustering of the cell types according to their hub promoters. The heat map shows the Jaccard score of each two cell types. (**G**) The Venn diagram of hub promoters derived from tB, tCD4 and Mon. (**H**) The Venn diagram of hub promoters derived from tCD4, tCD8 and nCD4.

keeping genes from (50) and detected a large fraction of overlaps between hub promoters and housekeeping genes, leading to a significantly high enrichment of hub promoters in housekeeping genes (*P*-value $= 1.27 \times 10^{-26}$, Fisher's exact test). This result partly explains why hub promoters are active across cell lines and also illustrates the biological meaning of hub promoters.

Fourth, we integrated a PPI network with 74 791 physical interactions derived from BIOGRID (26), HPRD (27) and MINT (28) databases to check the interaction degrees of proteins coded by hub promoters. The result shows that the average interaction degree of hub promoters is 13.31. To assess the significance of this average degree, we randomly generated protein groups at the same sample size for 100 000 times and then calculated the average degrees of these random groups, yielding a distribution of average degrees (Figure 5C). We found proteins encoded by hub promoters had a significantly higher average degree compared

Downloaded from https://academic.oup.com/nar/article-abstract/47/10/e60/5380496 by SCD - Universite Toulouse III user on 10 October 2019

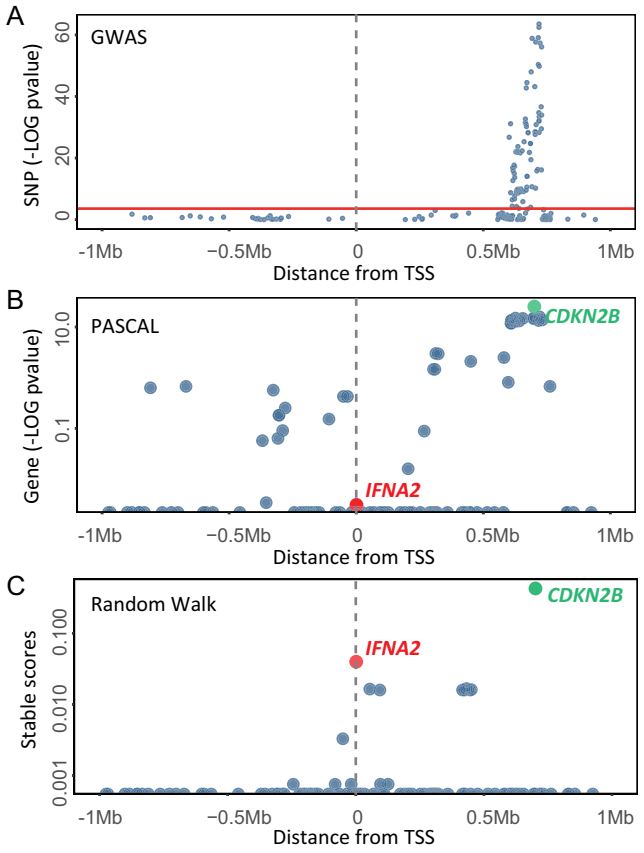

Figure 6. Manhattan plots of signals around *IFNA2*. (**A**) The Manhattan plot of GWAS SNP signals around the TSS of *IFNA2* (1 Mb upstream and 1 Mb downstream). The red horizontal line represents the genome-wide significance level (Bonferroni correction P-value $= 2.15 \times 10^{-4}$). (**B**) The Manhattan plot of initial PASCAL P values of regulatory elements around the TSS of *IFNA2*. (**C**) The Manhattan plot of stable scores of regulatory elements around the TSS of *IFNA2*. *CDKN2B* (green) and *IFNA2* (red) were ranked first and second by the random walk.

with random protein groups (P-value $= 1.30 \times 10^{-4}$), indicating proteins encoded by hub promoters also tend to be hubs in the PPI network.

Finally, we assessed the functional enrichment of hub promoters in GO terms (38) and REACTOME pathways (37). Results show that hub promoters are significantly enriched in the core biological processes and pathways (Figure 5D and E). For example, hub promoters show high enrichment in RNA metabolic processes, DNA binding events and organelle parts, suggesting interactions mediated by hub promoters play a vital role in the molecular functions and fundamental cell processes. Besides, hub promoters are highly enriched in cell cycle, promoter opening, chromosome maintenance and other chromatin-related pathways, implying hub promoters are closely associated with chromatin structures and promoter communication and regulation.

In summary, hub promoters defined by DeepTACT interactions are characterized by distinct features. To draw a more general conclusion, we analyzed the hub promoters defined by DeepTACT interactions detected in the other five cell types and obtained similar results (Supplementary Figures S8–S12). We further observed the hub promoters of different cell types and found a significantly large fraction of overlaps between each two groups of hub promoters (Figure 5F; Jaccard scores ranging from 0.291 to 0.467; P values $< 2.2 \times 10^{-16}$, Fisher's exact tests), supporting the point that hub promoters are fundamental across cell lines. Interestingly, we observed more overlaps among hub promoters of different types of the same cell line (i.e. tCD4, tCD8 and nCD4) than hub promoters of different cell lines (Figure 5G–H), indicating that hub promoters can reflect cell similarity. With this understanding, we applied hierarchical clustering to the six cell types based on Jaccard scores of their hub promoters (Figure 5F). The clustering result reveals the lineage relationship of different cell types, which is totally consistent with the hematopoietic tree (18).

Identification of disease-related regulatory elements using predicted interactions

We designed a computational method to identify disease-related regulatory elements by integrating summary statistics of genome-wide association study (GWAS) data of a given disease and chromatin contacts of a cell line related to the disease. We illustrate this by an initial example. Since there have been reports of the involvement of T cells in coronary artery disease (CAD) (51,52), we explore the use of interactions detected in total CD4+ T cells to identify CAD-related promoters and enhancers (application to tCD8 and nCD4 is shown in Supplementary Text S8, Tables S10 and S11). First, we collected 79 128 SNPs from the meta-analysis of GWA study including a total of 22 233 patients and 64 762 normal individuals (53). Meanwhile, we merged all interactions inferred by DeepTACT from total CD4+ T cells, as well as positive training interactions, into a highly sparse gene–enhancer network of 22 702 nodes and 40 993 edges. After excluding genes not coding proteins and cross-chromosome interactions, we simulated a random walk process on the gene–enhancer network with P values of nodes derived from GWAS data as initial probabilities, yielding a steady probability score for each node. The steady scores can serve as a measurement of the association between a gene/enhancer and the disease (see 'Materials and methods' section for details).

We noticed that regulatory elements ranked top by the random walk included not only those with high initial probabilities, but also those with insignificant P values (Supplementary Table S12). Since results of the random walk partly rely on the initial probabilities of nodes that are derived from GWAS data, it is not surprising to see nodes with highly significant P values ranked top. For example, *CDKN2B* (cyclin dependent kinase inhibitor 2B) ranks first with the most significant P-value (initial P-value $= 1.08 \times 10^{-13}$). *PSRC1* (proline and serine rich coiled-coil 1) ranks fifth by the random walk (initial P-value $= 5.19 \times 10^{-10}$). These genes are well known to be associated with CAD (54,55).

It is interesting to see *IFNA2* (interferon α2), which cannot be detected based on the GWAS signal (initial P-value $= 1$), ranks second by random walk. Specifically, we found that there was no significant SNP around the TSS of *IFNA2* (the nearest signal is more than 500 kb away; Figure 6A)

Downloaded from https://academic.oup.com/nar/article-abstract/47/10/e60/5380496 by SCD - Universite Toulouse III user on 10 October 2019

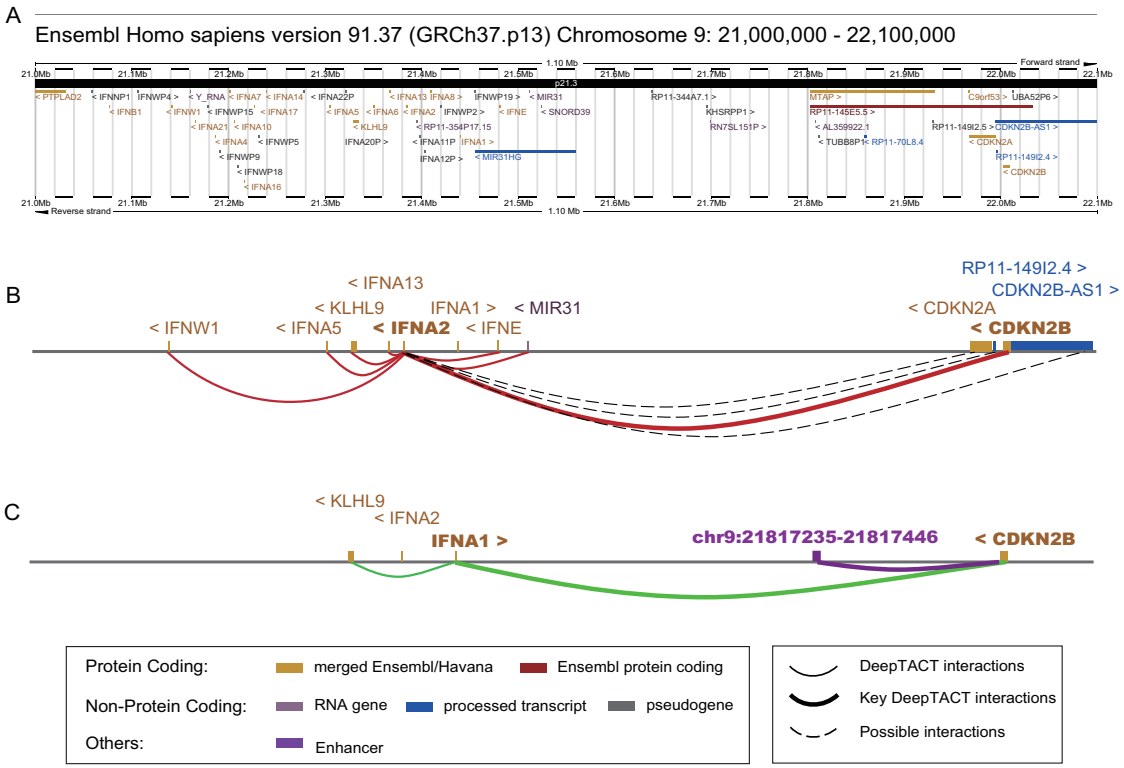

Figure 7. Fine-mapping interactions linking disease-related genes/enhancers. (**A**) Chromosome 9: 21,000,000 - 22,100,000 from Ensembl *Homo sapiens* version 91.37. (**B**) 3D contacts linking *IFNA2* to other genes (red curves). (**C**) 3D contacts linking *IFNA1* and chr9:21817235-21817446 to other genes (green and purple curves). Interactions between regulatory elements are displayed with full curves, while possible interactions derived from bait-level Hi-C data are represented by gray dashed curves.

and the initial probability of *IFNA2* was lower than 40 other regulatory elements within 2 Mb region (Figure 6B), making it hard to associate *IFNA2* with CAD based on the GWAS signals only. Intriguingly, when integrated with the gene–enhancer network, *IFNA2* was given the second highest steady score (Figure 6C), which implies a strong relationship between *IFNA2* and CAD. Indeed, it has been reported that *IFNA2* was significantly down-regulated in CAD patients (56). Meanwhile, we found our gene–enhancer network links *IFNA2* to *CDKN2B* (Figure 7A and B), a gene locates at a risk locus of CAD in 9p21.3 (57). Studies have shown that *IFNA2* and *CDKN2B* are both p53-mediated immunity genes (58) and are reported to have functional association related to CAD (59). This again supports our finding that *IFNA2* and *CAKN2B* are CAD-related genes. Besides *IFNA2*, we also identified two novel candidates for CAD, a gene *IFNA1* (interferon α1) and an enhancer chr9:21817235-21817446 (Figure 7C), which have not been reported to be related to CAD yet.

Altogether, this example suggests that the joint analysis of GWAS and the gene–enhancer network is helpful in prioritizing disease-related regulatory elements. To further highlight the contribution of the network constructed using interactions predicted by DeepTACT, we developed two other control networks to detect disease-related regulatory elements: one is a network constructed directly using candidate interactions derived from PCHi-C data, another is a network constructed using co-opening interactions selected from candidate pairs (details in Supplementary Text S9). Similarly, we conducted the random walk strategy separately on these two networks to detect disease-related regulatory elements. Results show that regulatory elements detected after integrating control networks are the same as those detected based on GWAS *P* values (Supplementary Tables S13 and S14), suggesting that the control networks did not provide enough additional information during the random walking process. In contrast, the integration of GWAS data and interactions predicted by our DeepTACT model succeeded in detecting meaningful disease genes, indicating that interactions predicted by DeepTACT can provide more information for the detection of disease-related regulatory elements than co-opening interactions and candidate interactions.

DISCUSSION

In this paper, we proposed a bootstrapping deep learning model, named DeepTACT, to predict 3D interactions between regulatory elements using their sequences and chromatin opening signals as input. In our work, we utilized DNase-seq data to provide chromatin accessibility information, while ATAC-seq data can also be used in the same way (Supplementary Text S10 and Figure S13). The large amount of public DNase-seq and ATAC-seq experiments makes the application of DeepTACT easier than other methods that are strict in model inputs (8–10). Based on the understanding that there exists a number of enhancer-like promoters that can regulate other promoters just like en-

Downloaded from https://academic.oup.com/nar/article-abstract/47/10/e60/5380496 by SCD - Universite Toulouse III user on 10 October 2019

hancers do (11,12), we paid equal attention to the detection and analysis of promoter–promoter interactions and that of promoter–enhancer interactions. Briefly, we first statistically demonstrated the ability of DeepTACT in identifying 3D interactions between regulatory elements. Then, we applied DeepTACT to PCHi-C datasets and showed that DeepTACT can perdict fine-grain interactions from PCHi-C data. We also defined a class of hub promoters and illustrated that these hub promoters were characterized by distinctive biological features. In addition, we elucidated how interactions between regulatory elements can be exploited for the identification of disease genes and the interpretation of disease mechanisms.

Several directions are worth exploring in the future. First, features learned by the deep learning model could be further explored to explain the relationship among the TFs enriched in different ends of interaction pairs. Although we have already reported a number of tissue-specific TFs learned by convolution kernels of DeepTACT from different tissues, it is hard to decipher the interactions between TFs using our present model, which has only one convolution layer for each regulatory element. A specially designed model is needed to answer this question. Second, in this work we applied DeepTACT to predict interactions from a series of promoter capture Hi-C data. We expect DeepTACT to be applied to general Hi-C data in future work. In the supplementary, we give an example of how DeepTACT can be applied to Hi-C data (Supplementary Text S11). Third, there are other choices to attain tissue-specific epigenomic information for this model, such as ChIP-seq of histone marks and TFs, and data for DNA methylation. A comparison of model performance given different types of epigenomic data as input will shed lights on the understanding of the relationship between different epigenomic events and chromatin contacts. Fourth, chromatin contacts connect genes to their regulators and thus help interpret regulatory effects on the expression level of target genes (Supplementary Figure S14). This interpretation makes interactions between regulatory elements useful in predicting gene expression. Finally, regulatory interactions can be used to score the influence of GWAS variants on the regulation mechanisms. Given a GWAS variant falling into any end of an interaction, we can assign a score for the variant based on the difference in predicted interaction probabilities between original sequences and sequences after mutation. In this way, interactions between regulatory elements can offer an opportunity for the annotation and interpretation of the non-coding genome.

DATA AVAILABILITY

Source code for DeepTACT model training is freely available at https://github.com/liwenran/DeepTACT.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Qijin Yin and Tianyi Sun for their helpful discussion about deep learning. We thank Mengmeng Wu for his useful suggestion on GWAS analysis. R.J. is a RONG professor at the Institute for Data Science, Tsinghua University. *Authors' Contributions*: R.J., W.H.W. and W.L. developed the concepts and designed the study. R.J. and W.H.W. instructed the whole research. W.L. designed the methods, performed all analyses and wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

National Key Research and Development Program of China [2018YFC0910404]; National Natural Science Foundation of China [61721003, 61873141, 61573207]; National Institutes of Health grants [R01HG007834, P50HG007735]; Tsinghua-Fuzhou Institute for Data Technology. Funding for open access charge: National Key Research and Development Program of China [2018YFC0910404]; National Natural Science Foundation of China [61721003, 61873141, 61573207]; National Institutes of Health grants [R01HG007834, P50HG007735].
Conflict of interest statement. None declared.

REFERENCES

1. Tang,R., Noh,H.J., Wang,D., Sigurdsson,S., Swofford,R., Perloski,M., Duxbury,M., Patterson,E.E., Albright,J. and Castelhano,M. (2014) Candidate genes and functional noncoding variants identified in a canine model of obsessive-compulsive disorder. *Genome Biol.*, **15**, R25.
2. Heidari,N., Phanstiel,D.H., He,C., Grubert,F., Jahanbani,F., Kasowski,M., Zhang,M.Q. and Snyder,M.P. (2014) Genome-wide map of regulatory interactions in the human genome. *Genome Res.*, **24**, 1905–1917.
3. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.
4. Fullwood,M.J., Liu,M.H., Pan,Y.F., Liu,J., Xu,H., Mohamed,Y.B., Orlov,Y.L., Velkov,S., Ho,A. and Mei,P.H. (2009) An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, **462**, 58–64.
5. Rao,S.S., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D. and Lander,E.S. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
6. Mifsud,B., Tavares-Cadete,F., Young,A.N., Sugar,R., Schoenfelder,S., Ferreira,L., Wingett,S.W., Andrews,S., Grey,W. and Ewels,P.A. (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.*, **47**, 598–606.
7. Zhang,Y., An,L., Xu,J., Zhang,B., Zheng,W.J., Hu,M., Tang,J. and Yue,F. (2018) Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. *Nat. Commun.*, **9**, 750.
8. Zhu,Y., Chen,Z., Zhang,K., Wang,M., Medovoy,D., Whitaker,J.W., Ding,B., Li,N., Zheng,L. and Wang,W. (2016) Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.*, **7**, 10812.
9. Al Bkhetan,Z. and Plewczynski,D. (2018) Three-dimensional epigenome statistical Model: Genome-wide chromatin looping prediction. *Sci. Rep.*, **8**, 5217.
10. Whalen,S., Truty,R.M. and Pollard,K.S. (2016) Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.*, **48**, 488–496.
11. Diao,Y., Fang,R., Li,B., Meng,Z., Yu,J., Qiu,Y., Lin,K.C., Huang,H., Liu,T. and Marina,R.J. (2017) A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods*, **14**, 629–635.
12. Gasperini,M., Findlay,G.M., McKenna,A., Milbank,J.H., Lee,C., Zhang,M.D., Cusanovich,D.A. and Shendure,J. (2017) CRISPR/Cas9-Mediated scanning for regulatory elements required for HPRT1 expression via thousands of large, programmed genomic deletions. *Am. J. Human Genet.*, **101**, 192–205.

Downloaded from https://academic.oup.com/nar/article-abstract/47/10/e60/5380496 by SCD - Universite Toulouse III user on 10 October 2019

13. Sun,Y., Wang,X. and Tang,X. (2014) *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, pp.1891–1898.
14. Szegedy,C., Vanhoucke,V., Ioffe,S., Shlens,J. and Wojna,Z. (2016) *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas,, pp. 2818–2826.
15. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
16. Kelley,D.R., Snoek,J. and Rinn,J.L. (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*, **26**, 990–999.
17. Park,Y. and Kellis,M. (2015) Deep learning for regulatory genomics. *Nat. Biotechnol.*, **33**, 825–826.
18. Javierre,B.M., Burren,O.S., Wilder,S.P., Kreuzhuber,R., Hill,S.M., Sewitz,S., Cairns,J., Wingett,S.W., Várnai,C. and Thiecke,M.J. (2016) Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*, **167**, 1369–1384.
19. Cairns,J., Freire-Pritchett,P., Wingett,S.W., Várnai,C., Dimond,A., Plagnol,V., Zerbino,D., Schoenfelder,S., Javierre,B.-M. and Osborne,C. (2016) CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol.*, **17**, 127.
20. Consortium,E.P. (2004) The ENCODE (ENCyclopedia of DNA elements) project. *Science*, **306**, 636–640.
21. Andersson,R., Gebhard,C., Miguel-Escalada,I., Hoof,I., Bornholdt,J., Boyd,M., Chen,Y., Zhao,X., Schmidl,C. and Suzuki,T. (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455–461.
22. Cunningham,F., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G. and Fitzgerald,S. (2014) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–D669.
23. Tang,Z., Luo,O.J., Li,X., Zheng,M., Zhu,J.J., Szalaj,P., Trzaskoma,P., Magalska,A., Wlodarczyk,J. and Ruszczycki,B. (2015) CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell*, **163**, 1611–1627.
24. Li,G., Chen,Y., Snyder,M.P. and Zhang,M.Q. (2017) ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis. *Nucleic Acids Res.*, **45**, e4.
25. Westra,H.-J., Peters,M.J., Esko,T., Yaghootkar,H., Schurmann,C., Kettunen,J., Christiansen,M.W., Fairfax,B.P., Schramm,K. and Powell,J.E. (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.*, **45**, 1238–1243.
26. Chatr-Aryamontri,A., Breitkreutz,B.-J., Oughtred,R., Boucher,L., Heinicke,S., Chen,D., Stark,C., Breitkreutz,A., Kolas,N. and O'donnell,L. (2014) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, **43**, D470–D478.
27. Keshava Prasad,T., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B. and Venugopal,A. (2008) Human protein reference database—2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
28. Chatr-Aryamontri,A., Ceol,A., Palazzi,L.M., Nardelli,G., Schneider,M.V., Castagnoli,L. and Cesareni,G. (2006) MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, **35**, D572–D574.
29. Wallace,B.C., Small,K., Brodley,C.E. and Trikalinos,T.A. (2011) *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, pp. 754–763.
30. Duren,Z., Chen,X., Jiang,R., Wang,Y. and Wong,W.H. (2017) Modeling gene regulation from paired expression and chromatin accessibility data. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, E4914–E4923.
31. Graves,A., Jaitly,N. and Mohamed,A.-R. (2013) *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, pp. 273–278.
32. Lamparter,D., Marbach,D., Rueedi,R., Kutalik,Z. and Bergmann,S. (2016) Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Comput. Biol.*, **12**, e1004714.
33. Jiang,R. (2015) Walking on multiple disease-gene networks to prioritize candidate genes. *J. Mol. Cell Biol.*, **7**, 214–230.
34. Li,W., Wang,M., Sun,J., Wang,Y. and Jiang,R. (2017) Gene co-opening network deciphers gene functional relationships. *Mol. Biosyst.*, **13**, 2428–2439.
35. D'Alessio,A.C., Fan,Z.P., Wert,K.J., Baranov,P., Cohen,M.A., Saini,J.S., Cohick,E., Charniga,C., Dadon,D. and Hannett,N.M. (2015) A systematic approach to identify candidate transcription factors that control cell identity. *Stem Cell Rep.*, **5**, 763–775.
36. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S. and Tokimatsu,T. (2007) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
37. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G. and Jassal,B. (2010) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
38. Consortium,G.O. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
39. Dao,L.T., Galindo-Albarrán,A.O., Castro-Mondragon,J.A., Andrieu-Soler,C., Medina-Rivera,A., Souaid,C., Charbonnier,G., Griffon,A., Vanhille,L. and Stephen,T. (2017) Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat. Genet.*, **49**, 1073–1081.
40. Ong,C.-T. and Corces,V.G. (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.*, **12**, 283–293.
41. Consortium,E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
42. Ganeshan,K. and Chawla,A. (2014) Metabolic regulation of immune responses. *Annu. Rev. Immunol.*, **32**, 609–634.
43. Osborn,O. and Olefsky,J.M. (2012) The cellular and signaling networks linking the immune system and metabolism in disease. *Nat. Med.*, **18**, 363–374.
44. Sanyal,A., Lajoie,B.R., Jain,G. and Dekker,J. (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109–113.
45. Koch,C.M., Andrews,R.M., Flicek,P., Dillon,S.C., Karaöz,U., Clelland,G.K., Wilcox,S., Beare,D.M., Fowler,J.C. and Couttet,P. (2007) The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res.*, **17**, 691–707.
46. Creyghton,M.P., Cheng,A.W., Welstead,G.G., Kooistra,T., Carey,B.W., Steine,E.J., Hanna,J., Lodato,M.A., Frampton,G.M. and Sharp,P.A. (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 21931–21936.
47. Benevolenskaya,E.V. (2007) Histone H3K4 demethylases are essential in development and differentiation. *Biochem. Cell. Biol.*, **85**, 435–443.
48. Barski,A., Cuddapah,S., Cui,K., Roh,T.-Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
49. Eisenberg,E. and Levanon,E.Y. (2003) Human housekeeping genes are compact. *Trends Genet.*, **19**, 362–365.
50. Eisenberg,E. and Levanon,E.Y. (2013) Human housekeeping genes, revisited. *Trends Genet.*, **29**, 569–574.
51. Sandberg,W.J., Yndestad,A., Øie,E., Smith,C., Ueland,T., Ovchinnikova,O., Robertson,A.-K.L., Müller,F., Semb,A.G. and Scholz,H. (2006) Enhanced T-cell expression of RANK ligand in acute coronary syndrome. *Arterioscler. Thromb. Vasc. Biol.*, **26**, 857–863.
52. Burren,O.S., García,A.R., Javierre,B.-M., Rainbow,D.B., Cairns,J., Cooper,N.J., Lambourne,J.J., Schofield,E., Dopico,X.C. and Ferreira,R.C. (2017) Chromosome contacts in activated T cells identify autoimmune disease candidate genes. *Genome Biol.*, **18**, 165.
53. Schunkert,H., König,I.R., Kathiresan,S., Reilly,M.P., Assimes,T.L., Holm,H., Preuss,M., Stewart,A.F., Barbalic,M. and Gieger,C. (2011) Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.*, **43**, 333–338.
54. Kral,B.G., Mathias,R.A., Suktitipat,B., Ruczinski,I., Vaidya,D., Yanek,L.R., Quyyumi,A.A., Patel,R.S., Zafari,A.M. and Vaccarino,V. (2011) A common variant in the CDKN2B gene on chromosome 9p21 protects against coronary artery disease in Americans of African ancestry. *J. Hum. Genet.*, **56**, 224–229.
55. Ripatti,S., Tikkanen,E., Orho-Melander,M., Havulinna,A.S., Silander,K., Sharma,A., Guiducci,C., Perola,M., Jula,A. and Sinisalo,J. (2010) A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet North Am. Ed.*, **376**, 1393–1400.
56. Jha,H.C., Srivastava,P., Vardhan,H., Singh,L.C., Bhengraj,A.R., Prasad,J. and Mittal,A. (2011) Chlamydia pneumoniae heat shock protein 60 is associated with apoptotic signaling pathway in human

Downloaded from https://academic.oup.com/nar/article-abstract/47/10/e60/5380496 by SCD - Universite Toulouse III user on 10 October 2019

atheromatous plaques of coronary artery disease patients. *J. Cardiol.*, **58**, 216–225.
57. Harismendy,O., Notani,D., Song,X., Rahim,N.G., Tanasa,B., Heintzman,N., Ren,B., Fu,X.-D., Topol,E.J. and Rosenfeld,M.G. (2011) 9p21 DNA variants associated with coronary artery disease impair interferon-[ggr] signalling response. *Nature*, **470**, 264–268.
58. Chaum,E., Winborn,C.S. and Bhattacharya,S. (2015) Genomic regulation of senescence and innate immunity signaling in the retinal pigment epithelium. *Mamm. Genome*, **26**, 210–221.
59. Vangala,R.K., Ravindran,V., Kamath,K., Rao,V.S. and Sridhara,H. (2013) Novel network biomarkers profile based coronary artery disease risk stratification in Asian Indians. *Adv. Biomed. Res.*, **2**, 59.

Downloaded from https://academic.oup.com/nar/article-abstract/47/10/e60/5380496 by SCD - Universite Toulouse III user on 10 October 2019